УДК 546:378.26(076)

СТАТИСТИЧЕСКИЙ АНАЛИЗ КАЧЕСТВА ТЕСТОВ, ПРИМЕНЯЕМЫХ ДЛЯ КОНТРОЛЯ ЗНАНИЙ ПО ХИМИИ

М.Г. Минин, Н.Ф. Стась, Е.В. Жидкова, О.Б. Родкевич

Томский политехнический университет
E-mail: minin@tpu.ru

С помощью математического аппарата классической теории тестов проведен анализ результатов экзаменов по химии. Вычислены статистические показатели первого и двух последних экзаменов. Показано приближение частотного распределения тестовых баллов к нормальному распределению статистических данных. Показано, что надёжность тестов находится в допустимом интервале значений, но необходимо увеличение их содержательной валидности.

Тестовые технологии контроля знаний и умений студентов широко применяются в вузах России на промежуточных этапах учебного процесса, но на итоговом контроле тестирование применяется редко. Применение тестирования на экзаменах сдерживается отсутствием данных о качестве применяемых тестов и надёжности получаемых результатов контроля.

В Томском политехническом университете разработана и используется тестовая технология контроля знаний студентов по химии, как на рубежном, так и на итоговом контроле студентов [1–3]. Создание научно-обоснованного теста состоит из четырёх этапов [4]: 1) планирование, 2) составление предтестовых заданий, 3) проведение апробационного тестирования, 4) коррекция заданий. Первые три этапа нами пройдены. Задача этой работы – исследование экспериментальных результатов тестирования для проведения работ по коррекции заданий теста. Актуальность работы объясняется необходимостью использования в высшем профессиональном образовании взаимопризнаваемых методов контроля обученности студентов, как это предусмотрено Болонской декларацией, к которой в 2003 г. присоединилась Россия.

Экзамен по тестовой технологии сдают студенты общетехнических направлений и специальностей, общее число проэкзаменованных составляет 2219 чел. Результаты экзамена, которые обрабатываются с помощью компьютера, являются исходным экспериментальным материалом. Он представлен в виде матрицы, число строк в которой соответствует числу испытуемых, а число столбцов – числу заданий в тесте. Фрагмент матрицы результатов последнего экзамена представлен в табл. 1. Каждое из 12 заданий теста в матрице пронумеровано дважды. Это связано с тем, что в наших тестах используются двухуровневые задания, которые позволяют контролировать не только конечные результаты, но и промежуточные этапы умственных действий студентов. Результаты выполнения оцениваются дихотомически: за правильный ответ студент получает 1 балл, а за неправильный или пропуск подзадания – 0.

В матрице отсутствуют строки и столбцы, состоящие только из нулей или только из единиц. Это означает, что среди студентов нет ни одного, который не выполнил все без исключения задания теста, и также нет ни одного, который выполнил правильно все задания. Соответственно нет ни одного задания, которое не выполнили или выполнили все студенты. По этим данным можно сделать предварительный вывод о том, что тесты более или менее сбалансированы по трудности заданий, а выборка испытуемых – репрезентативная.

Таблица 1. *Фрагмент матрицы результатов экзамена по химии в зимней экзаменационной сессии 2005/06 уч. г.*

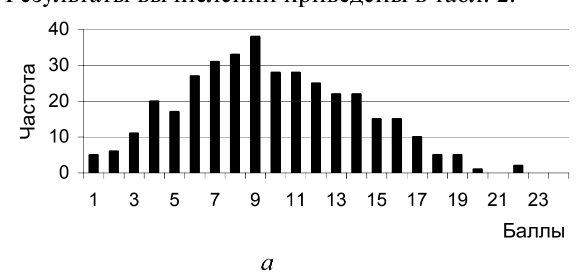
№ студента в списке	Задания																							
	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11	11	12	12
1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	0	0
3	1	1	1	1	1	1	0	0	1	1	0	1	1	1	0	0	0	0	1	1	1	1	1	0
4	1	1	0	1	1	1	1	0	1	0	1	1	1	1	0	0	1	0	0	0	1	1	1	1
5	1	0	1	1	0	1	0	1	0	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1
6	0	1	0	0	1	1	0	1	1	1	1	0	0	1	0	1	1	1	1	1	0	0	1	1
7	1	1	1	0	1	1	1	1	0	0	1	0	0	1	1	1	0	0	1	0	0	0	0	1
8	1	1	1	1	1	1	0	0	1	0	0	0	1	1	0	1	0	0	0	0	1	1	1	1
9	1	1	1	1	1	0	1	1	0	0	0	0	1	1	0	0	1	1	0	0	1	1	1	0
10	0	0	1	1	1	0	1	0	1	0	1	0	1	0	1	1	1	0	0	0	1	1	1	1
11	0	0	1	1	1	0	1	1	1	0	1	0	1	0	1	0	1	0	0	0	1	1	1	0
12	1	1	1	0	1	0	0	0	1	1	0	1	1	1	0	0	0	0	0	0	1	1	1	1

По результатам экзамена подсчитаны индивидуальные баллы каждого студента, число правильных ответов всех студентов на каждое задание и проведена графическая интерпретация полученных данных наиболее наглядным способом: в виде столбиковых гистограмм несгруппированных баллов. Гистограмма представляет собой последовательность столбиков, каждый из которых соответствует определенному результату экзамена (экзаменационному баллу), а высота столбца пропорциональна частоте «присутствия» этого балла в результатах экзамена. Иначе говоря, гистограмма представляет частотное распределение статистических результатов тестирования.

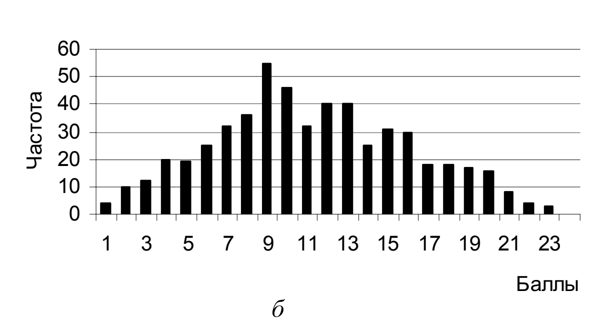
На рисунке приведены гистограммы результатов трёх экзаменов, проведенных по данной технологии – первого и двух последних. Из их сравнения видно постепенное улучшение результатов тестирования: число более высоких баллов увеличивается.

Следует обратить внимание на то, что частотное распределение соответствует обычному распределению статистических данных, но отличается от теоретического (нормального) распределения смещением результатов в сторону меньших баллов. Особенно заметно такое смещение на гистограмме первого экзамена. Такая ситуация в общем случае интерпретируется как несоответствие тестов уровню знаний испытуемых: тесты для данной выборки студентов обладают повышенной трудностью. Но преподаватели химии, которые разрабатывают задания, не намерены их упрощать, т. к. считают, что обучать и контролировать студентов необходимо на высоком уровне интеллектуальной трудности, что является основным принципом развивающего обучения. Между сессиями проводилась корректировка заданий: устранялись неопределённость и неоднозначность формулировок, приводилась к единообразию терминология и символика, устранялись стилистические погрешности. Но никаких замен трудных заданий на лёгкие при этом не допускалось.

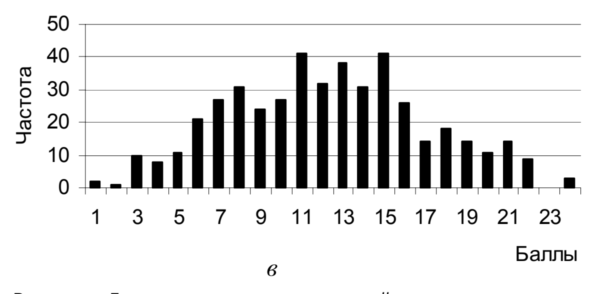
По результатам реального распределения частот методами математической статистики [4, 5] вычислены характеристики тестов, которые позволяют оценивать их в качестве инструмента измерения знаний: мода, медиана, среднее арифметическое, дисперсия, стандартное отклонение, асимметрия, эксцесс и самый важный показатель – надежность. Результаты вычислений приведены в табл. 2.



а



б



в

Рисунок. *Гистограммы результатов трёх экзаменов: зимнего 2003/04 (а), весеннего 2004/05 (б) и зимнего 2005/06 (в) учебного года*

Таблица 2. *Статистические характеристики тестов, применяемых для контроля знаний и умений студентов по химии*

Показатели	Зимняя сессия 2003/04 уч. г.	Весенняя сессия 2004/05 уч. г.	Зимняя сессия 2005/06 уч. г.
Число испытуемых	366	541	454
Мода	9	9	11, 15
Медиана	9	11	12
Среднее арифметическое	9,7	11,3	12,3
Дисперсия	17,6	23,4	22,6
Стандартное отклонение	4,2	4,8	4,8
Асимметрия	–0,23	–0,16	–0,09
Эксцесс	–0,40	–0,58	–0,52
Надёжность	0,71	0,75	0,76

Мода – это такое значение тестового балла, которое встречается наиболее часто среди результатов экзамена; в нашем случае оно равно 9 на экзаменах зимой 2004 и весной 2005 г. (унимодальное распределение). На последнем экзамене значений моды два: 11 и 15 (бимодальное распределение результатов); при этом второе значение моды заметно выше того, которым характеризуются результаты двух предыдущих экзаменов.

Медиана – это такое значение тестового балла, которое делит всех студентов на две равные части: с меньшим и большим значением результата экзамена; из табл. 2 видно, что этот показатель от сессии к сессии постепенно увеличивается.

Среднее арифметическое индивидуальных оценок экзамена, вычисленное обычным путем (все оценки суммируются с последующим делением на их число), равно 9,7; 11,3 и 12,3 балла. Отличие значения среднего арифметического от значения моды объясняется тем, что на величину первого влияют значения всех результатов, тогда как значение моды от других результатов не зависит. Увеличение значения среднего арифметического положительно характеризует процесс совершенствования методики преподавания дисциплины.

На практике значения моды, медианы и среднего арифметического следует иметь в виду при переводе тестовых баллов в традиционную оценку в том случае, когда результат экзамена (зачета) рассматривается как окончательный независимо от рейтинга студента в семестре. Такой подход применяется в кредитно-модульной системе, по которой обучаются студенты электротехнического института и факультета автоматики и вычислительной техники, и на которую в дальнейшем будут переходить другие подразделения Томского политехнического университета и других вузов.

Дисперсия и **стандартное отклонение** – показатели изменчивости (разброса) результатов тестирования. Дисперсия играет важную роль при разработке нормативно-ориентированных тестов: низкая дисперсия свидетельствует о слабой дифференциации тестируемых по уровню их подготовки, а высокая дисперсия приводит к большому отличию

получаемого распределения от теоретической нормальной кривой. Оптимальной считается такая дисперсия, при которой значение среднего арифметического равно утроенному значению стандартного отклонения. В наших тестах оптимального значения дисперсии пока не достигнуто, но тенденция положительна: если в первой сессии отношение среднего арифметического к стандартному отклонению составляет 2,3, то в последней – 2,6.

Асимметрия – показатель отклонения при тестировании распределения тестовых баллов от симметричного распределения, характерного для нормальной кривой. Асимметрия положительна, если большая часть тестируемых получает высокие оценки и отрицательна, если результаты тестирования противоположны. Положительная асимметрия характерна для облегченных тестов, отрицательная – для излишне трудных, а в хорошо сбалансированных по трудности тестах распределение баллов имеет вид нормальной кривой, для которой асимметрия равна нулю. На всех экзаменах мы наблюдаем отрицательную асимметрию (-0,23; -0,16; -0,09), которая постепенно уменьшается. Таким образом, наблюдается движение реального распределения тестовых баллов к теоретическому (нормальному) распределению, причем, происходит это не за счет снижения трудности заданий, а за счет их корректировки на соответствие тестологическим требованиям.

Эксцесс характеризует форму кривой распределения тестовых баллов, которая может быть островершинной (положительный эксцесс), средневершинной (нулевой эксцесс), характерной для нормальной кривой, и плосковершинной (отрицательный эксцесс). Из табл. 2 видно, что в нашем случае эксцесс имеет небольшое отрицательное значение, что свидетельствует о недостаточном числе студентов, получивших на экзаменах баллы, близкие к среднему значению.

Общая оценка статистических характеристик тестов такова: они приближаются к показателям нормального распределения результатов тестирования. Следовательно, появляется возможность шкалирования результатов тестирования, т. е. использования разрабатываемых тестов в качестве инструмента измерения знаний. Но если тест используется как инструмент измерения знаний, то он должен соответствовать определенным требованиям надежности и валидности.

Надежностью называется такая характеристика теста, которая показывает, насколько точны его измерения и насколько устойчивы результаты измерения к действию случайных факторов. Следовательно, тест надежен, если он обеспечивает высокую точность измерений и если результаты измерений устойчивы к действию внешних факторов (места и времени тестирования, состава тестируемых и т. д.).

Любое измерение содержит ошибки [4]: промахи, систематические ошибки и случайные ошибки. Промахи возникают при грубых нарушениях процедуры тестирования, но при наличии качественной инструкции и опыта проведения экзаменов они невозможны. Систематические ошибки возникают постоянно, поэтому они проявляют себя и могут быть устранены. В нашем случае такой ошибкой были неверные эталоны ответов в некоторых заданиях, которые легко были выявлены и исправлены. Случайные ошибки непредсказуемы и от них зависит точность измерений и надежность теста как инструмента измерения.

Теория надежности является важнейшей частью классической теории тестов. На практике используются три основных метода оценки надежности тестов: 1) повторное тестирование, 2) распределение группы, 3) расщепление теста.

По первому методу сравниваются результаты двух тестирований с помощью одного и того же теста с интервалом 2–3 недели одних и тех же испытуемых, когда они не успели забыть материал теста и не усваивали новые знания. На практике этот метод применяется редко, т. к. уровень знаний тестируемых в промежутке между тестированиями конечно же меняется. По второму методу все тестируемые разделяются на две группы, которые работают с одним и тем же тестом; результаты групп сравниваются: если результаты одинаковы или близки, то тест надежен.

Мы использовали третий метод, по которому сравниваются результаты выполнения двух частей теста. Разделений теста на две части может быть множество. При всех возможных разделениях теста коэффициент надежности вычисляется по формуле Кьюдера-Ричардсона [5]. Нижним допустимым значением коэффициента надежности, вычисленным по этой формуле, является значение 0,7. При более низком значении использование теста нецелесообразно из-за большой погрешности измерения.

Надёжность наших тестов, вследствие корректировки заданий на соответствие тестологическим требованиям, постепенно возрастает от значения 0,71 в первой сессии до 0,75 и 0,76 в двух последних экзаменационных сессиях. Достигнутая надёжность приемлема в практике тестирования, но она может быть более высокой. В этой связи необходимо знать все факторы, от которых зависит надежность тестов [6].

Длина теста. Надежность теста возрастает с увеличением его длины (числа заданий в нём), но увеличение длины предполагает не увеличение содержания, а детализацию проверки каждого элемента содержания дисциплины. Если установлена надежность при одной длине теста, то можно вычислить, насколько следует увеличить длину теста, чтобы повысить надежность до определенного значения. В нашем случае коэффициентом надежности 0,76 обладают тесты с числом заданий 24. Для повышения надежности до 0,80 длину теста следует довести до 30. Эта возможность не исключена, т. к. с одной стороны, числом заданий 24 проверяется лишь часть элементов содержания химии и, с другой стороны, три часа работы над действующи-

ми тестами – это большое время: большинство студентов выполняют тест за 2 часа.

Непонятность и двусмысленность заданий. Ответы на такие задания даются случайным образом, поэтому надежность измерения снижается. Составители заданий и эксперты должны изучать задания, поставив себя на место студентов, и исключать из тестов непонятные и двусмысленные задания.

Случайное угадывание правильных ответов. Возможность угадывания правильного ответа является самым распространенным поводом для критики тестов. Но эта проблема относится только к закрытым заданиям, при этом вероятность угадывания верного ответа уменьшается с увеличением числа дистракторов (отвлекающих ответов). При выборе ответа из двух альтернатив вероятность угадывания 50 %, при трех ответах – 33 %, при четырех – 25 % и т. д. Поэтому, чтобы уменьшить вероятность угадывания, составители закрытых заданий увеличивают число дистракторов. Но при этом нередко теряется чувство меры, и появляются некорректные дистракторы. В тестологии существуют формулы, по которым можно скорректировать оценку экзаменуемого, сделав поправку на вероятность случайного угадывания ответов. Но поправка снижает первичную оценку, что может быть причиной апелляции со стороны тех, кто не использует прием угадывания. Кардинальное решение проблемы угадывания, по которому мы идём, – уменьшение числа заданий закрытой формы.

Субъективное оценивание. Оно возможно при проверке результатов выполнения теста людьми, но исключается при компьютерном тестировании, при введении ответов студентов в компьютер независимыми наблюдателями, а также шифрованием экзаменационных работ. Последние два метода применяются в нашей технологии проведения экзамена.

Ошибка в подсчетах. В нашей работе она исключается использованием компьютера и специальной программы обработки результатов тестирования.

Качество инструкций. Инструкция должна быть понятной для всех экзаменуемых. Правила представления результатов выполнения заданий, оговоренные в инструкции, должны быть максимально простыми. В инструкции для того, кто проводит экзамен, должны быть четко определены правила его «поведения»; он не должен что-то подсказывать студентам или отвлекать их внимание. Присутствие лектора и преподавателей на экзамене нежелательно.

Состояние экзаменуемых и условия проведения экзамена могут повлиять на надежность результатов тестирования. Поэтому экзамен нежелательно проводить сразу после праздника или выходного дня. Необходимо проверить состояние аудитории (температура, освещенность, шумность, запахи и т. д.) и нейтрализовать факторы, снижающие надежность тестирования.

Ещё одной важной характеристикой тестов является их валидность.

Валидностью называется характеристика способности теста служить поставленной цели измерения. Существует несколько видов валидности: содержательная, диагностическая, прогностическая и т. д. При итоговом контроле знаний с использованием критериально-ориентированных тестов на первое место выступает содержательная валидность. Если тест позволяет проверить все то, что задумано авторами, то он является валидным относительно контролируемого содержания дисциплины.

В некоторых работах, например в [7], рассматривают валидность каждого тестового задания, оценивая его величиной точечно-бисериального коэффициента корреляции между результатами ответа тестируемых на данное задание и их суммарным и индивидуальными баллами. Конечно, каждый исследователь имеет право на свою точку зрения, но общепринято считать, что содержательная валидность является характеристикой теста, а не тестового задания.

Содержательная валидность определяется экспертным методом. Преподаватели химии считают, что действующими тестами проверяются не все знания и умения студентов, что содержательная валидность экзаменационных тестов должна быть более высокой. Поэтому актуальна работа по составлению новых заданий, соответствующих тем элементам содержания, на которые нет или недостаточно имеющихся тестовых заданий. Кроме того, необходим расчёт и анализ характеристик тестовых заданий с целью выявления и замены тех, которые выходят за рамки разумных требований по трудности и дискриминативности.

Выводы

1. Математическая обработка результатов экзаменационных сессий по химии, проведенных по тестовой технологии, показала постепенное улучшение характеристик используемых тестов.
2. Частотное распределение результатов двух последних экзаменов приближается к теоретическому (нормальному) распределению статистических данных, поэтому появляется возможность использования тестов в качестве инструмента измерения знаний.
3. Надежность тестов (76 %) находится в допустимом интервале значений, но может быть повышена при дальнейшей целенаправленной работе над тестовыми заданиями.
4. Необходимы расчёт и анализ характеристик используемых тестовых заданий, а также составление новых заданий, направленных на увеличение содержательной валидности используемых тестов.

СПИСОК ЛИТЕРАТУРЫ

1. Минин М.Г., Стась Н.Ф., Жидкова Е.В., Родкевич О.Б. Тестовая технология контроля знаний студентов по химии // Известия Томского политехнического университета. – 2005. – Т. 308. – № 4. – С. 231–235.
2. Минин М.Г., Стась Н.Ф., Жидкова Е.В. Внутривузовская система диагностики качества знаний студентов // Качество высшего профессионального образования: достижения, проблемы, перспективы: Матер. Всеросс. научно-практ. конф. 21–23 января 2005 г. / Алтайский государственный технический университет им. И.И. Ползунова. – Барнаул: Изд-во АлтГТУ. – 2005. – С. 250–252.
3. Минин М.Г., Стась Н.Ф., Жидкова Е.В. Автоматизированный контроль знаний студентов технического вуза // Качество образования: менеджмент, достижения, проблемы: Матер. VI Межд. научно-метод. конф. 23–25 мая 2005 г / Новосибирский государственный технический университет. – Новосибирск: Изд-во НГТУ, 2005. – С. 321–324.
4. Челышкова М.Б. Теория и практика конструирования педагогических тестов. – М.: Логос, 2002. – 432 с.
5. Карпенко Д.С., Карпенко О.М., Шлихунова Е.Н. Автоматизированная система мониторинга эффективности усвоения знаний и качества тестовых заданий // Инновации в образовании. – 2001. – № 2. – С. 69–84.
6. Майоров А.Н. Теория и практика создания тестов для системы образования. – М.: Народное образование, 2000. – 352 с.
7. Стрельникова Е.Н. Анализ результатов централизованного тестирования по химии 2004 года // Вопросы тестирования в образовании. – 2004. – № 11. – С. 76–84.

Поступила 06.06.2006 г.

УДК 373.880